

Measuring contact patterns with wearable sensors: methods, data characteristics and applications to data-driven simulations of infectious diseases

A. Barrat^{1,2,3}, C. Cattuto³, A. E. Tozzi⁴, P. Vanhems^{5,6} and N. Voirin⁵

1) Aix Marseille Université, CNRS, CPT, UMR 7332, Marseille, 2) Université de Toulon, CNRS, CPT, UMR 7332, La Garde, France, 3) Data Science Laboratory, ISI Foundation, Torino, 4) Bambino Gesù Children's Hospital, IRCCS, Rome, Italy, 5) Hospices Civils de Lyon, Hôpital Edouard Herriot, Service d'Hygiène, Epidémiologie et Prévention, Equipe Epidémiologie et Biomarqueurs de l'Infection and 6) Laboratoire de Biométrie et de Biologie Evolutive, Université de Lyon, Université Lyon 1, CNRS UMR 5558, Equipe Epidémiologie et Santé Publique, Lyon, France

Abstract

Thanks to recent technological advances, measuring real-world interactions by the use of mobile devices and wearable sensors has become possible, allowing researchers to gather data on human social interactions in a variety of contexts with high spatial and temporal resolution. Empirical data describing contact networks have thus acquired a high level of detail that may yield new insights into the dynamics of infection transmission between individuals. At the same time, such data bring forth new challenges related to their statistical description and analysis, and to their use in mathematical models. In particular, the integration of highly detailed empirical data in computational frameworks designed to model the spread of infectious diseases raises the issue of assessing which representations of the raw data work best to inform the models. There is an emerging need to strike a balance between simplicity and detail in order to ensure both generalizability and accuracy of predictions. Here, we review recent work on the collection and analysis of highly detailed data on temporal networks of face-to-face human proximity, carried out in the context of the SocioPatterns collaboration. We discuss the various levels of coarse-graining that can be used to represent the data in order to inform models of infectious disease transmission. We also discuss several limitations of the data and future avenues for data collection and modelling efforts in the field of infectious diseases.

Keywords: Contact patterns, data-driven, infectious diseases, modelling, numerical simulations, wearable sensors

Article published online: 25 November 2013

Clin Microbiol Infect 2014; **20**: 10–16

Corresponding author: A. Barrat, CPT, Campus de Luminy, Case 907, 13288 Marseille cedex 9, France
E-mail: alain.barrat@gmail.com

Introduction

Contact patterns among individuals play an important role in determining the potential transmission routes of infectious diseases. Knowledge of these patterns is thus relevant for identifying contagion pathways, for informing models of epidemic spread, and for the design and evaluation of control measures, such as the targeting of specific groups of individuals with appropriate prevention strategies or interventions (e.g. drug prophylaxis, vaccination, hand-washing, and the use of masks) [1–5].

Empirical descriptions of contact patterns have, until recently, mostly relied on interviews and surveys, sometimes on a very large scale [6–13], yielding important insights. Surveys allow distinction between different types of contact (e.g. involving or not involving physical contact) and classification of the contacts according to their context (at work, at home, at school, etc.), and have yielded information on the mixing patterns of different age groups in various countries [10].

However, surveys are costly and often have often a low response [13], and the precise formulation of the question

might influence the answers. Answers are also subject to memory biases, which are difficult to estimate [12]. Moreover, surveys often collect ego-networks on single days (see, however, [12,14]), and it is then difficult to estimate some properties of the contact networks known to be relevant for the spread of infectious diseases, such as the number of triangles and the fraction of repeated contacts from one day to the next [15,16].

The use of novel technologies, in particular of networked wearable sensors, offers appealing alternatives: Bluetooth or Wi-Fi can be used to assess the proximity of individuals [17,18], and even the face-to-face presence of individuals can be resolved with high spatial and temporal resolution [19]. Here, we review recent work developed within the SocioPatterns collaboration (www.sociopatterns.org), where wearable proximity sensors were used to collect large-scale datasets on human face-to-face interactions in various contexts, including conferences, hospitals, schools, and museums [16,19–21].

Data Collection: Method, Statistical Analysis, and Representations

The data collection infrastructure is based on wearable wireless devices that exchange radio packets in a peer-to-peer fashion to monitor the location and proximity of individuals (www.sociopatterns.org). The use of ultra-low-power radio signals allows radio packets to be exchanged only between devices located within 1–1.5 m of one another. Moreover, when individuals wear the devices on their chest and the lowest radio power is used, exchange of packets between devices is only possible when they are facing each other, as the human body acts as a shield at the radio frequency used. In summary, the system detects and records close-range encounters during which a communicable disease infection could be

transmitted, e.g. by coughing, sneezing, or hand contact (Fig. 1 and [19]).

The sensing system is tuned so that the recorded data include, for each detected contact between participants, its start and end times, with a temporal resolution of approximately 20 s: it is thus possible to monitor the number of contacts that each individual establishes with any other individual, the duration of individual encounters, the cumulative time spent in contact between two or more individuals, the frequency of encounters, and how these quantities evolve during the observation period.

The complexity of the data is revealed through statistical analysis of the contact event durations and of the time intervals between contacts: large variability is indeed observed in all these quantities. The corresponding distributions, shown in Fig. 2a, b, are broad, similarly to other characteristics of human behaviour [22]: short durations are the most probable, but very long durations are also observed with a non-negligible probability, and no characteristic temporal scale emerges. For transmissible diseases for which the transmission probability between two individuals depends on their time in contact, this means that different contacts might yield very different transmission probabilities: many contacts are very short and correspond to a small transmission probability, but some are much longer than others, and could therefore play a crucial role in disease dynamics. These strong duration fluctuations constitute a robust property observed across all contexts [19–21] and at different moments, although the amount of activity varies significantly from case to case (Fig. 2c).

The detailed knowledge of all contact events allows the re-creation of an artificial *in silico* population and hence the simulation of potential spreading phenomena with a high degree of realism. It is also possible to test, in simulations, the impacts of specific interventions for mitigation and containment. Highly detailed contact data might, however, represent

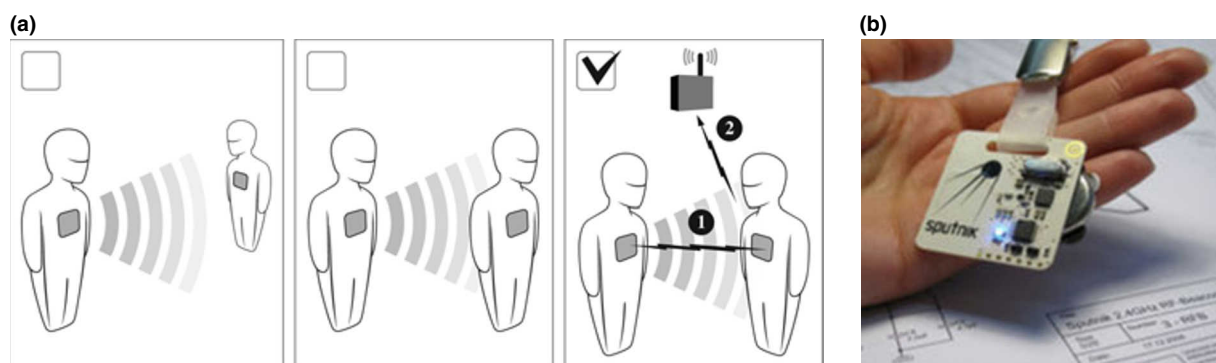


FIG. 1. (a) Schematic illustration of the sensing infrastructure. Radio frequency identification (RFID) devices are worn as badges by the individuals participating in the deployments. A face-to-face contact is detected when two persons are close and facing each other. The interaction signal is then sent to RFID readers located in the environment. (b) RFID device worn by participants.

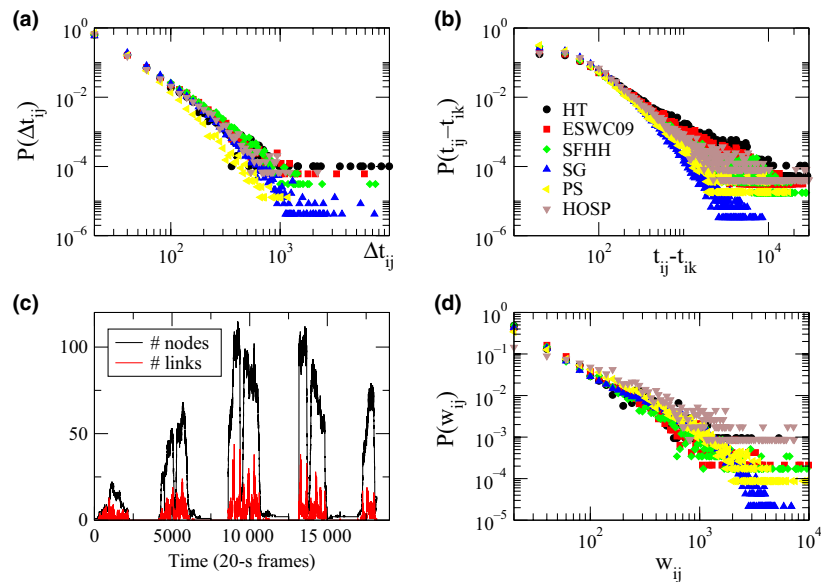


FIG. 2. Statistical properties of the contact data for several datasets (see Table I for the dataset characteristics). (a) Probability of observing a contact of duration Δt vs. Δt , computed as the number of contacts of duration Δt divided by the total number of contacts. (b) Probability of observing a time interval of a given duration between two successive contact events of a given individual, aggregated over the entire population. (c) Evolution of the number of nodes and links in 20-s instantaneous networks during a conference. (d) Probability of observing a daily cumulated contact duration w_{ij} between individuals i and j (i.e. number of pairs $i-j$ with daily cumulated contact duration w_{ij} , divided by the total number of pairs of individuals who have been in contact at least once during the day).

TABLE I. Partial list of the datasets on face-to-face proximity collected by the SocioPatterns collaboration during 2009 and 2010 and discussed in the present article

Name	Date	Venue	Event type	No. of participants	Duration
SG	April–July 2009	Science Gallery, Dublin, Ireland	Exhibition	c. 30 000	3 months
ESWC09	June 2009	ESWC 2009, Crete, Greece	Conference	c. 180	4 days
SFHH	June 2009	SFHH, Nice, France	Conference	c. 400	2 days
HT09	July 2009	ACM Hypertext 2009, Torino, Italy	Conference	c. 120	3 days
PS	October 2009	Primary school, Lyon, France	School	c. 250	2 days
OBG	November 2009	Bambino Gesù Hospital, Rome, Italy	Hospital	c. 100	10 days
HOSP	December 2010	Edouard Herriot Hospital, Lyon, France	Hospital	c. 100	4 days

unnecessarily detailed information if the goal is to simply extract stylized facts or generic statistics of contacts, which can be used to design and validate models of human contacts and to estimate the relevance of transmission control strategies. The observed sequence of contacts might indeed be influenced by specific aspects of the environment in which the measure took place, and represents only one instance of many contact sequences occurring in the same environment on different days. It is thus useful to build contact summary statistics, which are expected to be more robust to variations of the specific measurement context. To this end, the time-resolved contact data are usually aggregated along two dimensions, as discussed below.

First, aggregation along the temporal dimension yields cumulative contact networks preserving the information at

the individual level: such networks describe who has been in contact with whom, and each link between two nodes is weighted by the cumulative time spent in contact by the two corresponding individuals. The resulting weights are highly heterogeneous, as shown in Fig 2d): whereas most links correspond to very short durations, some correspond to very long contact times, with no characteristic interaction time-scale. Similar statistics have been observed in various contexts, ranging from scientific conferences to schools, hospitals, or offices [3,16,19–21,23]. The heterogeneity of contact patterns at the individual level is known to have a strong impact on spreading dynamics [24,25]. In particular, it highlights the existence of ‘super-contactors’, i.e. individuals who account for an important proportion of the overall contact durations and may therefore become super-spreaders in the case of an outbreak [23,26].

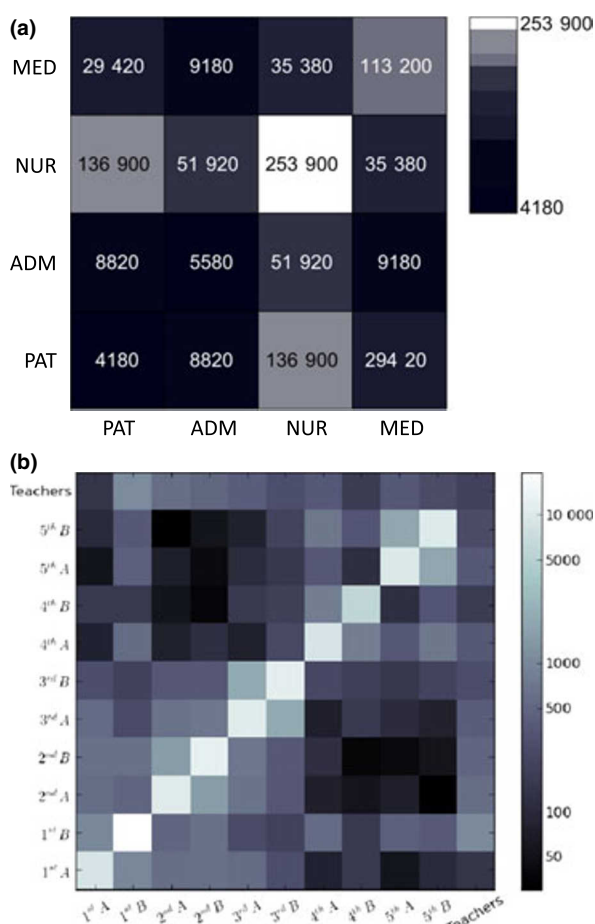


FIG. 3. Contact matrices giving the cumulated durations in seconds of the contacts between classes of individuals, measured in the HOSP and PS datasets. In the hospital case, individuals were categorized, according to their role in the ward, as nurses, doctors, patients, and administrative staff. In the school case, individuals were categorized according to the classes that they were in (here, ranging from 1st to 5th grade). The matrix entry at row X and column Y gives the total duration of all contacts between all individuals of class X with all individuals of class Y. NUR, paramedical staff (nurses and nurses' aides); PAT, patient; MED, medical doctor; ADM, administrative staff.

High-resolution contact data can also be aggregated over specific attributes of the individuals. When the population under study is structured, i.e. when individuals can be classified according to specific characteristics or role (e.g. according to their age class or professional activity), a convenient representation of their contacts is provided by contact matrices, whose elements give the number (or duration) of the contacts that individuals in one given class have with individuals of another class, as illustrated in Fig. 3. Such a representation is useful for designing interventions, as it can suggest easily generalizable strategies that target specific classes of individuals.

Contact matrices, however, typically report only averages, discarding the strong fluctuations in the numbers and durations of contacts between two individuals of given classes. The contact matrix representation also carries the implicit assumption that all individuals are in contact with one another: any two individuals are assumed to be connected, with a weight that depends only on their relative classes.

The cumulative duration of contacts between two individuals, however, fluctuates strongly, even when their role classes are fixed [21,27]: for instance, the contact duration between nurse A and doctor B can be very different from the one between nurse C and doctor D. Information on the the average contact duration between nurses and doctors is thus not sufficient. Moreover, the density of links connecting individuals in given classes depends on the specific classes, and is sometimes very low: many pairs of individuals never have any contact. In order to account for these important properties, the concept of 'contact matrix of distributions' (CMD) was introduced [27]: in this representation, as in usual contact matrices, the contact patterns between individuals depend on their relative classes. For each pair of classes, the distribution of cumulated contact durations is fitted to a certain functional form (e.g. a negative binomial in order to account for its broad character), and the matrix elements are given by the parameters of this fit. Therefore, the CMD representation does not retain the specific information on who has been in contact with whom, but it does retain both the empirical density of links and the heterogeneity of contact durations.

Finally, each of the above representations can be computed for the entire duration spanned by the dataset or for restricted time windows, such as half-days or days, in order to investigate possible variations with time of the contact statistics, networks, or matrices [20,23,27]. Information on temporal variations of individuals' contacts from one day to the next could be included in the contact matrix data summaries, e.g. as an additional parameter in each contact matrix entry (giving, for example, for nurses and doctors, the average fraction of new contacts between nurses and doctors from one day to the next).

Using High-resolution Contact Data in Models of Epidemic Spread

Empirical time-resolved contact data can be used to perform data-driven simulations of epidemic spread in a population [3,16,27]. As each dataset describes a specific population and environment, and therefore features specificities that might not be representative of another period or context, an issue naturally emerges: what level of detail on the contact patterns

should be incorporated into computational models of spread, so that the relevant information is retained but the model stays as parsimonious and generalizable as possible? In other words, what are the most useful synopses of high-resolution contact patterns? On the one hand, data summaries that are too coarse might disregard important properties; on the other hand, representations that are too fine might be too specific, and the integration of highly detailed data may yield models that are less transparent and lead to results that are less general in their applicability [16,27,28].

In order to shed light on these issues, we have investigated the impact of the data representation on simulations of epidemic spread by building a hierarchy of data representations corresponding to different levels of aggregation. Each representation was used to simulate the spread of an infectious disease [16,27], and the results were compared with the outcome of simulations based on the most detailed representation (regarded as a reference standard). This methodology provides several insights. First, the time at which the epidemic peaks is a robust property of the spread. It is correctly approximated even by simulations based on coarse data representations, at least for spreading processes that are slow with respect to the temporal resolution of the data [16,27]. Moreover, data representations that do not take into account the heterogeneity of contact durations lead to overestimations of the probability of a large outbreak and the attack rate (Fig. 4) [16,27]. Most importantly, they might lead to an incorrect classification of the relative risks for different classes of individuals [27]. In contrast, the representation by CMDs allows the correct modelling of important features of epidemic spread, and the estimation of the relative risks of individuals in different role classes, while maintaining a parsimonious form that retains very little information from the time-varying contact data that it summarizes. This representation thus provides a practical tool that translates complex properties of the contacts within a population into practically actionable information for guiding intervention and prevention strategies. It represents an interesting synopsis of highly detailed contact data that combines simplicity, compactness of representation, and modelling power, which are essential features for guiding decision-making in public health contexts.

Discussion

Infection control still represents one of the most important challenges in public health. The current policies for control of infections transmitted through person-to-person contact are based on general assumptions that may not be applicable to all individuals, meaning that the policies may be difficult to apply

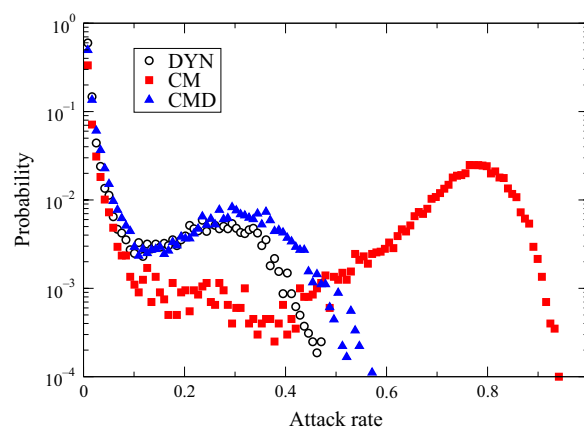


FIG. 4. Probability of observing a certain attack rate (fraction of the final number of cases) for numerical simulations of a susceptible–exposed–infected–recovered model on the OBG dataset (i.e. number of simulations leading to a given attack rate, divided by the total number of simulations). The different curves correspond to simulations performed with different representations of the raw data. DYN, dynamical contact network including the precise start and end time of each contact; CM, contact matrix representation that includes only information on the average duration of contact between individuals of different classes (individuals are categorized here as nurses, assistants, doctors, patients, and care-givers); CMD, contact matrix of distributions that takes into account the heterogeneity of contact durations and the different densities of links among different categories of individuals. Simulations performed with the CM representation lead to an overestimation of the final number of cases.

to an entire population. Strategies driven by detailed data on the contact patterns within populations promise to increase efficiency and feasibility. To this end, it is crucial to identify appropriate methods for using high-resolution data to inform models and design prevention strategies.

The unsupervised measurement of contact patterns with wearable sensors provides an interesting opportunity: it gives access to the network of contacts between individuals, and provides key information on the structure and heterogeneity of contacts between individuals belonging to different role classes and on the repetition of contact patterns on different days. The collection of data in diverse contexts highlights differences and similarities between human contacts, depending on context, and has shown the remarkable robustness of crucial statistical features such as the distribution of cumulated contact durations. The collected data can be used to design models of human behaviour, inform models of epidemic spread, and design and evaluate containment strategies in diverse contexts such as schools and hospitals. For instance, it is possible to evaluate, in simulations, the performance of targeted vaccination strategies or the role of the intervention

timing. It is also possible to envision novel intervention types, such as changes in the daily schedule of a school or in the organization of a hospital ward, or to estimate the relative efficiency of school closure and targeted class closure [29].

The present methodology has some limitations. Most datasets collected so far correspond to the contacts in populations of relatively limited size (a few hundred individuals) over a limited amount of time. No information is collected on contacts occurring either outside the range of the sensing system or involving individuals not wearing sensors. Moreover, the data do not provide information on physical contacts (unlike some surveys) or on the occurrence of events that are known to favour transmission, such as coughing and sneezing.

These limitations hint at future research directions. Further data collection campaigns will be crucial to validate and consolidate the results across other hospital units and other contexts, and over longer periods of time. Detailed comparisons with surveys [13] would also constitute an important cross-validation tool. The role of sampling of the population should also be carefully assessed.

Additional datasets will also help in the evaluation of the use of proxies, such as the ones put forward in [2,30], that may replace systematic detailed measurement of contact patterns. For example, the use of metadata on school schedules or shifts in hospitals allows us to infer approximate co-location properties, group structures and spatial distributions that can be used to simulate epidemic spread and design intervention strategies. High-resolution data from wearable sensors could then be used as a reference standard to validate the results of the simulations that use only proxy data.

High-resolution datasets can also be used to devise models of human mobility and contact that can be used to generate synthetic datasets that can be fed into models of disease spread at various scales, helping to eventually achieve multi-scale models that span several temporal, spatial and contact scales.

The overlay and integration of epidemiological, microbiological and genomic information with contact patterns among individuals may also enable radical changes in the approach to infection control. The precise measure of the contact pattern in a population could be combined with whole genome sequencing techniques for pathogens to investigate outbreaks. So far, only traditional surveys have been used to investigate social networks in these situations [31]. Combining phylogenetic analysis of viruses with actual contact data could provide valuable information about the transmission mechanisms of infectious diseases, especially regarding the role of frequency and/or duration of contacts. The inclusion of host susceptibility characteristics (e.g. age, sex, underlying

diseases, and genetic susceptibility) in such datasets would also enable more precise studies of infectious disease transmission mechanisms.

Acknowledgements

This work was partially supported by the French ANR project HarMS-flu (ANR-12-MONU-0018 to A. Barrat) and by the EU FET project Multiplex 317532 to A. Barrat and C. Cattuto. This study was also partially supported by the FINOVI foundation, the INSERM IMI Pandemic programme, and GOJO.

Transparency Declaration

The authors declare no conflicts of interest.

References

1. Chowell G, Viboud C. A practical method to target individuals for outbreak detection and control. *BMC Med* 2013; 11: 36.
2. Smieszek T, Salathe M. A low-cost method to assess the epidemiological importance of individuals in controlling infectious disease outbreaks. *BMC Med* 2013; 11: 35.
3. Salathe M, Kazandjieva M, Lee JW, Levis P, Feldman MW, Jones JH. A high-resolution human contact network for infectious disease transmission. *Proc Natl Acad Sci USA* 2010; 107: 22020–22025.
4. Cauchemez S, Bhattarai A, Marchbanks TL et al. Role of social networks in shaping disease transmission during a community outbreak of 2009 H1N1 pandemic influenza. *Proc Natl Acad Sci USA* 2011; 108: 2825–2830.
5. Temime L, Opatowski L, Pannet Y, Brun-Buisson C, Boelle PY, Guilletot D. Peripartetic health-care workers as potential superspreaders. *Proc Natl Acad Sci USA* 2009; 106: 18420–18425.
6. Read JM, Edmunds WJ, Riley S, Lessler J, Cummings DA. Close encounters of the infectious kind: methods to measure social mixing behaviour. *Epidemiol Infect* 2012; 140: 2117–2130.
7. Beutels P, Shkedy Z, Aerts M, Van Damme P. Social mixing patterns for transmission models of close contact infections: exploring self-evaluation and diary-based data collection through a web-based interface. *Epidemiol Infect* 2006; 134: 1158–1166.
8. McCaw JM, Forbes K, Nathan PM et al. Comparison of three methods for ascertainment of contact information relevant to respiratory pathogen transmission in encounter networks. *BMC Infect Dis* 2010; 10: 166.
9. Mikolajczyk RT, Akmatov MK, Rastin S, Kretzschmar M. Social contacts of school children and the transmission of respiratory-spread pathogens. *Epidemiol Infect* 2008; 136: 813–822.
10. Mossong J, Hens N, Jit M et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Med* 2008; 5: e74.
11. Zagheni E, Billari FC, Manfredi P, Melegaro A, Mossong J, Edmunds WJ. Using time-use data to parameterize models for the spread of close-contact infectious diseases. *Am J Epidemiol* 2008; 168: 1082–1090.
12. Smieszek T, Burri EU, Scherzinger R, Scholz RW. Collecting close-contact social mixing data with contact diaries: reporting errors and biases. *Epidemiol Infect* 2012; 140: 744–752.

13. Danon L, Read JM, House TA, Vernon MC, Keeling MJ. Social encounter networks: characterizing Great Britain. *Proc R Soc B* 2013; 280: 20131037.
14. Read JM, Eames KT, Edmunds WJ. Dynamic social networks and the implications for the spread of infectious disease. *J R Soc Interface* 2008; 5: 1001–1007.
15. Smieszek T, Fiebig L, Scholz RW. Models of epidemics: when contact repetition and clustering should be included. *Theor Biol Med Model* 2009; 6: 11.
16. Stehlé J, Voirin N, Barrat A et al. Simulation of an SEIR infectious disease model on the dynamic contact network of conference attendees. *BMC Med* 2011; 9: 87.
17. O'Neill E, Kostakos V, Kindberg T et al. Instrumenting the city: developing methods for observing and understanding the digital cityscape. *Lect Notes Comput Sci* (2006)4206: 315–332.
18. Pentland A. *Honest signals: how they shape our world*. Cambridge, MA: MIT Press, 2008.
19. Cattuto C, Van den Broeck W, Barrat A, Colizza V, Pinton JF, Vespignani A. Dynamics of person-to-person interactions from distributed RFID sensor networks. *PLoS One* 2010; 5: e11596.
20. Stehlé J, Voirin N, Barrat A et al. High-resolution measurements of face-to-face contact patterns in a primary school. *PLoS One* 2011; 6: e23176.
21. Isella L, Romano M, Barrat A et al. Close encounters in a pediatric ward: measuring face-to-face proximity and mixing patterns with wearable sensors. *PLoS One* 2011; 6: e17144.
22. Barabasi A-L. Bursts: the hidden pattern behind everything we do. Dutton Adult. (2010)
23. Vanhems P, Barrat A, Cattuto C et al. Estimating potential infection transmission routes in hospital wards using wearable proximity sensors. *PLoS One* 2013; 8: 73970.
24. Anderson RM, May RM. *Infectious diseases of humans: dynamics and control*. Oxford: Oxford University Press, 1992.
25. Pastor-Satorras R, Vespignani A. Epidemic spreading in scale-free networks. *Phys Rev Lett* 2001; 86: 3200–3203.
26. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. Superspreading and the effect of individual variation on disease emergence. *Nature* 2005; 438: 355–359.
27. Machens A, Gesualdo F, Rizzo C, Tozzi AE, Barrat A, Cattuto C. An infectious disease model on empirical networks of human contact: bridging the gap between dynamic network data and contact matrices. *BMC Infect Dis* 2013; 13: 185.
28. Blower S, Go MH. The importance of including dynamic social networks when modeling epidemics of airborne infections: does increasing complexity increase accuracy? *BMC Med* 2011; 9: 88.
29. Gemmetto V, Barrat A, Cattuto C Mitigation of infectious diseases at school: targeted class closures vs school closures, to be submitted.(2013)
30. Curtis DE, Hlady C, Pemmaraju SV, Polgreen P, Segre AM Modeling and estimating the spatial distribution of healthcare workers. *Proceedings of the 1st ACM International Health Informatics Symposium IHI'(2010)* 10: 287–296.
31. Gardy JL, Johnston JC, Ho Sui SJ et al. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *N Engl J Med* 2011; 364: 730–739.